Gov 2002: Problem Set 6

Problem Set Instructions

This problem set is due on **November 1, 11:59 pm** Eastern time. Please upload a PDF of your solutions to Gradescope. We will accept hand-written solutions but we strongly advise you to typeset your answers in Rmarkdown. Please list the names of other students you worked with on this problem set.

Question 1 (30pt)

This problem will use the `subprime` data to walk you through a very common inference problem - testing whether the difference between two population values is non-zero. To begin this problem first download `subprime.csv` and load it into R.

Hint for coding:

- To load a csv file in R, type `data <- read.csv(`*`name of the file`*`)`. Then you can use an object called `data` throughout your R code.
- To load libraries (i.e., packages), type `library(`*`name of the package`*`)`. In this problem set, we suggest that you load ggplot2 and dplyr. If you have not installed these packages, install them by `install.packages(`*`name of the package`*`)`.

We are going to be interested primarily in the `loan.amount` variable - the amount that each loan recipient received. Suppose a lawsuit has been filed in U.S. District Court by a group of Fort Myers women who claim that women in the area were loaned less money than men. The defendants – a group of local mortgage lenders – are vigorously denying these claims, and the case is now advancing to trial. Having heard about your expertise in this area, the federal judge hearing the case has brought you in to provide expert testimony. Your task in this problem is to assist the judge in her determination.

a. Suppose you were only able to interview 100 male and 100 female loan recipients at random, making them iid. To simulate this in R, set the seed to 02138 and draw 100 observations randomly from the male subset of the `subprime` data and 100 observations randomly from the female subset of the data. These 200 observations constitute your sample. Calculate (1) the average loan amount (`loan.amount`) for women in your sample, (2) the average loan amount for men, and (3) and (4), the sample standard deviation for each.

Hint for coding:

- To set the seed, type `set.seed(02138)`.

- To subset a dataframe called `data` based on a value (e.g., 0) of a variable called `var1`, type `data[data$var1==0]`.
- To generate 100 random sample IDs from a vector $c(1, ..., N)$ with replacement, type `sample(1:N, 100, T)`.
- To get the mean of a variable `var1` of a dataframe `data`, type `mean(data$var1)`. For SD, use `sd(data$var1)`.

b. Let μ_m and μ_w be the population average loan amount for men and women respectively. Let σ_m^2 and σ_w^2 be the population variances in loan amount for men and women respectively. Denote the sample average loan amounts for men by $\bar{X}_m$ and for women by $\bar{X}_w$. What is the expected value of the sampling distribution of $\bar{X}_m - \bar{X}_w$? What is the variance of the sampling distribution of $\bar{X}_m - \bar{X}_w$?

c. Compute and report your sample difference in average loan amount for men and women. Recall that for large samples, the sampling distribution of a mean or difference-in-means is approximately normal. Suppose that we know that the true population $\sigma_m^2 = 32381.57$ and $\sigma_w^2 = 19097.95$. Using the normal approximation, what is the probability that we would observe a difference-in-means at least as extreme as the one in our sample if the true population difference-in-means $\mu_m - \mu_w$ equals 0? Note that by "at least as extreme," we mean a value that is further away from 0 than the value we observe - that is, $P(|\bar{X}_m - \bar{X}_w| \geq \alpha)$ where α is our observed value and $||$ is the absolute value operator.

Hint:

- To get $P(X > x)$, consider calculating $1 - P(X < x)$.

Hint for coding:

- To get the probability that a normally distributed random variable takes on a value less than or equal to some value `q` using the command `pnorm(q, mean, sd)` where `mean` is the mean of the normal distribution and `sd` is the standard deviation.

d. Comment on your result in (c). Given what we observe in our sample, is it likely that there is no difference in loan amounts for men and women? A common threshold for "rejecting" our assumed hypothesis that $\mu_m - \mu_w = 0$ is observing a sample that would occur with probability .05 or less if that hypothesis were true (that is, a very unlikely sample). Would we reject the hypothesis that there is no difference in average loan amounts between men and women?

Question 2 (30pt)

In this problem, we will explore the implications of the Central Limit Theorem for uncertainty estimation and hypothesis testing. Start by creating two variables, `X1` and `X2`, using the following code:

```r
set.seed(02139)
X1 <- rnorm(100000, 5, 2)
X2 <- rexp(100000, 0.2)
```

For the purposes of this problem, we will treat these variables (each with 100,000 elements) as the full population. We will take samples from these two datasets to evaluate the coverage probability of 95% confidence intervals for the population mean using different types of data and different sample sizes.

a. Plot and describe the full distributions of `X1` and `X2`. What is the population mean and population standard deviation of each random variable (this is just the `mean()` and `sd()` of both of these variables)?

- **Hint for coding**:
  - Consider using the ggplot2 package (for example, `geom_density` would be helpful). There are various resources available on the web on how to use ggplot.

b. Now, create a loop to take 100 samples of size 8 from each dataset and record the sample mean for each sample. Plot the density of sample means for `X1` and `X2` separately and compare your these densities. Are they similar or different? Why?

- **Hint for coding:**
  - One way to work on this problem is to use for loop. If you want to loop over 1, 2, . . . , n, then your for loop starts by `for(i in 1:n) {...}`.
  - First create an empty matrix (or dataframe) to save results. Then save outputs from the for loop to that empty matrix.

c. Given your answer in part (a), what should be standard error of the sample mean of `X1` and the sample mean of `X2`? We know that if a random variable is normally distributed, 95% of its distribution will be within 1.96 standard deviations of the mean. What proportion of the sample means for `X1` and `X1` are within 2 standard errors of the population means?

d. Repeat the simulation in parts (b) and (c) for samples of size 8, 20, 50, and 500, and increase your number of simulations to 1000. Report the probability of being within 1.96 standard errors of the popoulation mean for each of your eight simulations in a table (you do not need to create additional plots). How do your results change? What differences do you see between `X1` and `X2`?

- **Hint for coding**:
  - One way to work on this problem is to loop over different sample sizes. Of course, you are free to use other functionalities if you wish.

e. Interpret your findings in parts (b)-(d). How does the central limit theorem explain what you see in the simulations?

Question 3 (20pt)

All probability distributions have *moments*, which are standard expressions that define its shape in ways you've already heard of and other more nuanced ways (the variance, the skew, kurtosis, etc.). Describing a population distribution (or empirical sample distribution) in terms of its moments is really useful in social science (e.g. the skew of income in the U.S. population is positive) Specifically, the nth central moment of a random variable X is defined as $E[(X - E[X])^n]$, but it is more common to work with the nth moment defined as as $E[X^n]$ (getting rid of the $E[X]^n$ part).

Suppose the random variable X for your population has the the following first four moments: $E[X] = 1/2$, $E[X^2] = 1/2$, $E[X^3] = 3/4$, $E[X^4] = 3/2$. Suppose you took an i.i.d. sample $\{X_1, \ldots, X_{20}\}$ of size 20 from this distribution. Let $T = (X_1^2 + ... + X_{20}^2)/20 = \overline{X^2}$, an estimator of the second moment.

a. What are $E[T]$ and $V(T)$? Be sure to explain why.

b. Use the central limit theorem to approximate the probability (in `R`) that T is less than or equal to 1. **Hint for coding:** Consider using `pnorm`.

Question 4 (20pt)

In class we learned that if a the variance of a sequence of random variables with finite mean goes to zero as $n \to \infty$, then the sequence will converge in probability to some value. But this is a sufficient condition, not a necessary one. To see this, consider the sequence of random variables X_n be with probability distribution:

$$X_n = \begin{cases} 0 & \text{with probability } 1 - 1/n \\ n & \text{with probability } 1/n \end{cases}$$

(a) Find $\mathbb{E}[X_n]$.

(b) Use the definition of convergence in probability to show that $X_n \xrightarrow{p} 0$.